TabDDPM: Modelling Tabular Data with Diffusion Models

Year: 2022 | Citations: 381 | Authors: Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, Artem Babenko

---

Abstract

Denoising diffusion probabilistic models are currently becoming the leading paradigm of generative modeling for many important data modalities. Being the most prevalent in the computer vision community, diffusion models have also recently gained some attention in other domains, including speech, NLP, and graph-like data. In this work, we investigate if the framework of diffusion models can be advantageous for general tabular problems, where datapoints are typically represented by vectors of heterogeneous features. The inherent heterogeneity of tabular data makes it quite challenging for accurate modeling, since the individual features can be of completely different nature, i.e., some of them can be continuous and some of them can be discrete. To address such data types, we introduce TabDDPM -- a diffusion model that can be universally applied to any tabular dataset and handles any type of feature. We extensively evaluate TabDDPM on a wide set of benchmarks and demonstrate its superiority over existing GAN/VAE alternatives, which is consistent with the advantage of diffusion models in other fields. Additionally, we show that TabDDPM is eligible for privacy-oriented setups, where the original datapoints cannot be publicly shared.